

DOCUMENT RESUME

ED 120 205

TM 005 164

AUTHOR Bernknopf, Stanley; Bashaw, W. L.
TITLE An Investigation of Criterion-Referenced Tests Under Different Conditions of Sample Variability and Item Homogeneity.
PUB DATE [Apr 76]
NOTE 27p.; Paper presented at the Annual Meeting of the American Educational Research Association (60th, San Francisco, California, April 19-23, 1976)
EDRS PRICE MF-\$0.83 HC-\$2.06 Plus Postage
DESCRIPTORS Comparative Analysis; *Criterion Referenced Tests; *Item Analysis; Language Tests; *Norm Referenced Tests; *Test Construction; *Test Reliability; Undergraduate Students
IDENTIFIERS Language Skills Examinations; Test Homogeneity; Test Theory; University System of Georgia; Variance (Statistical)

ABSTRACT

The present study was designed to examine whether or not traditional procedures concerning item selection and reliability are both applicable and appropriate for criterion-referenced (CR) tests. It was also designed to examine traditional procedures and those designed especially for CR testing in relation to test variance and item homogeneity. Specifically, the following questions were formulated: (1) How are traditional and criterion-referenced item selection techniques interrelated? (2) How are traditional and criterion-referenced reliability indices interrelated? (3) To what degree does variability among subjects and homogeneity of items affect traditional and criterion-referenced item selection procedures and reliability indices? The results of the study indicate that the construction of CR tests can be greatly facilitated by item analysis procedures such as phi, and the application of traditional reliability estimates such as KR-20. The procedure of trying out test items on a group of examinees consisting of master and nonmasters is recommended. (Author)

* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *

ED120205

SESSION
D-21

AN INVESTIGATION OF CRITERION-REFERENCED TESTS
UNDER DIFFERENT CONDITIONS OF SAMPLE VARIABILITY
AND ITEM HOMOGENEITY

Stanley Bernknopf
Pioneer Cooperative Educational
Service Agency
P.O. Box 548
Cleveland, Ga. 30528

and

W. L. Bashaw
University of Georgia
Fain Hall
Athens Ga. 30602

U.S. DEPARTMENT OF HEALTH,
EDUCATION & WELFARE
NATIONAL INSTITUTE OF
EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

Presented at the Annual Meeting of the
American Educational Research Association

2

San Francisco

April, 1976

TM005 164

AN INVESTIGATION OF CRITERION-REFERENCED TESTS
UNDER DIFFERENT CONDITIONS OF SAMPLE VARIABILITY
AND ITEM HOMOGENEITY

Stanley Bernknopf
Pioneer CESA

W. L. Bashaw
University of Georgia

The present study is designed to examine whether or not traditional procedures concerning item selection and reliability are both applicable and appropriate for criterion-referenced (CR) tests. It is also designed to examine traditional procedures and those designed especially for CR testing in relation to test variance and item homogeneity. Specifically, the following questions were formulated:

1. How are traditional and criterion-referenced item selection techniques interrelated?
2. How are traditional and criterion-referenced reliability indices interrelated?
3. How are traditional and criterion-referenced item selection techniques effected by subject variability and test homogeneity?
4. How are traditional and criterion-referenced referenced reliability indices effected by subject variability and test homogeneity?

Introduction

Within the last decade, the interest in mastery learning and accountability has led to renewed interest in CR testing. Along with this interest has come a surge of discussion concerning the appropriateness of using traditional test theory as a perspective from which to view CR measurement. Much of this discussion centers on the issues of item selection and reliability. Since CR test scores are believed to be restricted in range and therefore variance, a prerequisite requirement for classical test statistics, many researchers protest the use of such techniques.

Recently, item selection procedures and estimates of reliability for CR tests have been investigated using modified classical theory

approachs. The few investigations which have attempted to evaluate these procedures have done so in limited situations and have not yielded definite conclusions. In essence, these investigations have posed more questions than they have answered.

During the late 60's and early 70's a considerable number of articles on this issue appeared. In essence, two points of view have surfaced. On one side we have the feeling that since NR and CR instruments are used for different purposes, the usual item analysis procedures and reliability coefficients have limited if any use for CR tests (Roudabush, 1973). CR tests are generally intended to be prescriptive, while NR tests are generally intended to be descriptive. The other side professes that the difference between NR and CR tests stems from differences in how the respective scores are interpreted, and therefore classical test theory is both appropriate and desirable for CR tests (Haladyna, 1974).

Those researchers who argue against the use of classical test theory for CR measurement center their argument around the issue of variance. The traditional approaches to test construction and reliability rely heavily on the presence of response variance in test scores.

Where this variance is low, as would be the case when a particularly effective instructional treatment resulted in homogeneously high learner performance, the range restriction distorts the meaning of these correlationally derived estimates (Popham, 1974; p.11).

There is no question that lack of test score variance would attenuate item discrimination indexes and reliability estimates. However, there is some question as to whether or not the lack of test score variance is more a theoretical than an actual problem. The reality of the situation may simply be that students do vary in their performance. Reports of no variance may in essence "stem from failure to sample enough students and/or from the failure to examine the rate at which students master items and objectives" (Klein & Kosecoff, 1973; p.8). Woodson (1974) and Haladyna (1974) elaborate on this point when they hypothesize that restriction in range of CR test scores may be due to the selection of examinees rather than some intrinsic and unique characteristic of CR tests.

Popham and Husek (1969) have argued that variability is not a necessary condition for a good CR test. Woodson (1974), addressing himself to this point, feels that Popham and Husek fail to consider the question of what generalizations are to be made from the observations. Woodson feels that the "population of observations for which an item is calibrated must be the reference for the evaluation of an item...and not the sample statistics" (p.63). Referring to the situation where instruction is perfect, "the range of possible observations of interest includes the observation of passing no items as well as that of passing all items" (p.63). Woodson goes on to say that items in a CR test must be able to discriminate within the range of interest on the characteristic being measured. Item analysis in this case, requires observations at different points on the characteristics, perhaps pre-post instruction. Woodson concludes his article by stating that "item variance and discrimination are essential...items and tests which give no variability...give no information and are therefore not useful" (p.64).

Millman and Popham (1974) published a rejoinder to Woodson's article in which they again maintained that "item or score variability is an unnecessary characteristic of criterion-referenced tests as they have been traditionally conceived" (p.137). In essence, Millman and Popham feel that the primary purpose of a CR test is to determine how well an examinee performed in relation to some domain of tasks of which the test items are only a sample. In order to accomplish this, there is no reason to believe that the instrument must have item and test score variability. The criteria employed for selecting a test should be whether or not the test reflects the examinee's status with respect to the domain. Woodson(1974) responded to Millman and Popham by emphasizing that the basis for selecting items is and should be determined by their intended purpose, whether it be criterion-referenced or norm-referenced. "To argue that variance of items is essential is not the same as to argue for a particular procedure for selection of items which tend to maximize variance in a particular population of persons" (p.139).

Concerning an appropriate procedure for item selection, Haladyna (1974) believes that "samples consisting of both mastery and nonmastery

examinees offer a better discription of the degree of reliability possessed by...CR tests" (p.8). Results reported by Haladyna showed that samples based on post instruction alone yielded homogeneity estimates ranging from .31 to .72. However, when unrestricted samples were created by combining pre-and post-instruction test scores, homogeneity increased to a range of .69 to .89 with a median of .84. In addition to investigating the relationship between sample variance and homogeneity estimates, Haladyna also examined the relationship between the number of subscales and full scale homogeneity. In this case, he found no relationship between the number of subscales and the full scale homogeneity estimate. Haladyna reported that "...the subscales representing these regions may be highly internally consistent...and the full scale homogeneity may also be reasonably high" (p.8). Based on this, Haladyna sees classical test theory very applicable and desirable when dealing with the internal consistency of CR tests, especially when these tests have numerous regions or are multiscale tests.

Methodology

The term CRT as it applies to the present study is defined as a test designed to provide information on specific knowledge or skills possessed by a student. Furthermore, the test should provide information as to whether or not the student has reached competency. The instrument used in this study, The Language Skills Exam (LSE), conforms to this definition. The LSE, designed by the University System of Georgia, measures the competency of college students in the areas of reading and writing. The exam was developed specifically to provide state-wide information on the status of student competence in these areas and to provide a uniform means of identifying those students who fail to attain minimum levels of competence (Rentz, 1973). The LSE consists of three components: (1) an objective writing section, (2) a reading comprehension section, and (3) an essay writing sample. Components one and two are further divided into sub-parts.

Objective writing component Component one consists of seventy multiple-choice items divided into two parts. The item content in part 1 includes diction, sentence structure, point of view, spelling, punctuation, usage, and vocabulary. The thirty items in this part

consist of students selecting from among four alternatives the words or phrases most appropriate for inclusion in the stem. For each item, the stem consists of a sentence with a word or phrase deleted. Part 2 of the objective writing component contains forty items dealing with editing errors in usage, spelling, punctuation, and capitalization. The items require that the student identify the part of the sentence containing the error or indicate that no error exists.

Reading Comprehension Component The reading comprehension component contains items representing five content categories: (1) reproducing ideas, (2) identifying ideas and making inferences, (3) analyzing motivation, (4) analyzing presentation, and (5) criticizing a selection. During this part, the subject is presented with nine reading sections. Each selection is followed by five multiple-choice items, totaling forty-five items.

Essay Writing Component The essay component consists of having the student write an essay based on an assigned topic. The essays are graded according to a model essay comparative method. The method is based on a four point scale where "1" represents sub-standard performance and "4" represents high quality. Three model essays are used, one representing each borderline case on the four point scale. Each paper is graded either above or below a model. A model designated "3/2" distinguishes between a rating of "2" and "3". A paper worse than "3/2" would be rated "2" if it were better than the "2/1" model. A paper better than the "3/2" but worse than the "4/3" model would be rated "3". The following diagram illustrates the rating scale.

Ratings	4	3	2	1
Model designation	4/3	3/2	2/1	

The difference between a rating of "1" and a rating of "2" is the difference between the subject being able to reach minimum competency. The grade of "2" is assigned to those papers which meet only the basic criteria (Thompson and Rentz, 1973).

It is important to note that the LSE was designed as a multi-scale test with all three subtests measuring different aspects of the same domain, language skills. However, even though components one and two are highly intercorrelated, as separate subtests they represent a higher degree of item homogeneity than they do as a combined test.

The form of the LSE used in the present study differed from traditional forms in that it contained 53 experimental items. These items were being tried out for the purpose of obtaining item statistics. Students across the 31 institutions were randomly chosen to receive this form. This form as administered consisted of 70 writing items of which 33 were experimental. The reading component consisted of 40 items of which 20 were experimental. The essay component remained unchanged.

In order to prevent the relative size of the writing and reading component from influencing the variables under investigation, the following item manipulations were performed: (1) the 70 item writing component of the test was reduced to 40 items consisting of the 33 experimental plus seven of the non-experimental items; (2) The reading component remained in its original form; (3) when the reading and writing components were examined as one test, every other item was chosen so as to keep the test length constant at 40 items and to assure that there were 20 writing and 20 reading items.

Population and Sample Selection:

The population used in this study consisted of 1955 students in the 31 member institutions of the University System of Georgia who took the special form of the LSE in the Spring quarter of the 1973-74 school year. Students are eligible to take the LSE after they have completed between 60 and 75 quarter hours of college credit.

Three specific samples were drawn in such a manner as to create high, middle and low achievement groups with identical distributions and variances. In addition, an unrestricted sample consisting of subjects across achievement levels was drawn. Each of the samples, restricted and unrestricted, consisted of 400 subjects. Achievement on the writing skills component was used to select subjects corresponding to the desired samples.

The selection procedure first entailed obtaining a total population score distribution for the writing skills component of the test. This distribution consisted of scores which ranged from 0 to 31, with a mean of 18.23 and a standard deviation of 5.52. From this distribution, three score ranges were chosen to correspond to the three achievement levels desired for the restricted samples. Each range consisted of nine score categories which would allow for the

formation of a normal distribution based on stanines. The three distributions overlapped by five stanine points. Based on the appropriate percentages, subjects within each score category were randomly chosen until the desired number of subjects were reached. The unrestricted sample was chosen in a similar manner. For this sample the distribution was allowed to extend from the lowest score category of the low achievement sample to the highest score category of the high achievement sample. As with the restricted samples, subjects were randomly selected based on stanines. Percentages and actual number of subjects chosen are presented in Table 1.

By employing the above procedure along with the item manipulations discussed earlier, the following conditions were created: (1) the shape and the standard deviations of the three restricted distributions were held constant; (2) the number of items for the writing and reading components were held constant when considered as separate tests or as one test.

Test Homogeneity:

In addition to achievement level and sample variability, test homogeneity was also manipulated in the present study. Just as the three achievement levels when considered independently represented a lesser degree of variability than when considered together, the writing and reading components when considered as separate tests should represent a greater degree of homogeneity than when considered as one 40 item test (20 items from each component).

As previously mentioned, the sampling process used was designed to create four achievement samples. Three of these were to represent low, middle, and high achievement while the fourth was allowed to extend across the other three creating an unrestricted sample. Table 2 summarizes the results of t-tests between all combinations of sample means as well as F-ratios between sample variance. It should be noted from this table that while the difference between sample means are quite small in some cases, all differences except one are statistically significant. The one exception is between the middle achievement and the unrestricted samples. This is not surprising since the unrestricted sample, which contains low, middle, and high achievement subjects, would be expected to have a mean very close to the middle achievement sample. For both the reading and the writing and reading component,

the comparison of low achievement vs. middle achievement, and middle achievement vs. unrestricted achievement produced non-significant F-ratios between the variances. In addition, for the reading component, low achievement vs. high achievement also yielded a non-significant F-ratio between variances.

Analysis

Three sets of item selection indices were computed for the reading component alone and for the reading and writing component combined: (1) the traditional item selection procedure of employing the point biserial coefficient using total test score as a criterion (r_t); (2) the phi coefficient obtained in the fourfold table for each item, where the dichotomies consist of marking the item correct or incorrect on the test and of obtaining a total score above and below the point designated as mastery on the essay test (ϕ); (3) the point biserial as a correlation between the item and score on the essay test, an independent criterion as opposed to total test score (r_e).

For each set of items selected in the above manner, indices of internal consistency and reliability were calculated for the best 2/3 of the items. These indices consisted of: (1) The Kuder-Richardson coefficient (KR-20). (2) Livingston's reliability coefficient (r_{cc}) which redefines the variance as a deviation from the mastery criterion score rather than from the mean, and (3) Harris' index (uc^2) based on between and within group sums of squares that are determined by the two groups resulting from the dichotomization into mastery and non-mastery categories. The coefficient is equivalent to the squared Pearson product-moment correlation between total score on the test and the dummy variable designating the sort. In essence, it is a squared point biserial correlation coefficient.

These statistics were calculated for the unrestricted achievement sample as well as for the three restricted achievement samples and for the reading component alone as well as the reading and writing component combined. These manipulations resulted in the following groupings: (1) total sample and total test (high subject variability, low item homogeneity); (2) total sample with reading component (high subject variability, high item homogeneity); (3) for each achievement level, the total test (low subject variability, low item homogeneity); (4) for each achievement level, the reading component (low subject

variability, high item homogeneity).

The above analysis allowed the researcher to compare traditional and criterion-referenced item selection methods, to compare traditional and criterion-referenced reliability indices, and to study the relationship between them. These comparisons were made under differing conditions of sample variability and item homogeneity. Figure 1 summarizes the above comparisons.

Results and Discussion

Item selection

Table 3 presents the Spearman rank order correlation coefficients between item selection techniques for each achievement sample and test situation.

The overall trend is toward agreement. As was expected, the correlations between ϕ and r_e are higher than when either one is correlated with r_t . This was anticipated since both ϕ and r_e used the essay component as a criterion for selecting items while r_t used the total test score. In all but two cases, the relationship between the traditional index, r_t , and either of the two nontraditional indices, ϕ and r_e is statistically significant and in the direction of agreement. The low achievement reading component data produced correlations which were not significant when r_t was compared with either ϕ or r_e .

Within each achievement sample, the correlations between the three indices were consistently higher when the item pool consisted of the writing and reading components combined as opposed to the reading component alone. This would seem to indicate that, irrespective of sample ability, we would expect to get a greater proportion of items selected in common by the three indices as the item pool becomes more heterogeneous by covering a greater number of objectives.

Table 4 shows the percentages of common items that would appear in a final test when the test length is a proportion of the total item pool. Two test lengths are presented; one third of the item pool and two thirds of the item pool. In addition, the percentages of items discarded in common in the bottom one third are also presented. As will be noted from this table, the percentages increase considerably when the best two-thirds is considered as compared to the best one-third.

In order to further examine the differences and similarities between the three selection methods, internal consistency estimates were computed for different tests consisting of the best two thirds of the items produced by each method. Table 5 presents this data by sample and test situations. For each sample and test situation, the range between the highest and lowest mean produced by the three methods was quite small. In only one case, high achievement with reading alone, did the range exceed one raw score point.

The traditional point biserial and the two CR item selection techniques seem to produce tests containing many of the same items. As a result, the tests yield similar means and internal consistency coefficients. Initial inspection of the data would lead one to conclude that there is little difference between the item selection methods investigated. However, it should be pointed out that some items which did not contribute to the discrimination between mastery and nonmastery groups, as determined by ϕ and r_e , were retained when r_t was used to evaluate the items. Similarly, some of the best discriminations between mastery and nonmastery groups were made by items which would have been eliminated by using the conventional point biserial. For example, three items from the low achievement reading component data ranked 3rd, 8th and 10th in discriminating ability when using the conventional point biserial. The same items ranked 32nd, 36th and 30th, respectively, when ϕ was used. In this example, these items would have been retained using the conventional method and discarded using ϕ . In yet another case, item 40 from the unrestricted sample with writing and reading combined ranked 13th in discriminating ability when using the conventional point biserial, but ranked 33rd when mastery was used as the criterion. There are several other cases in which items would have been retained using one method and discarded using another. These cases were not limited to any one achievement sample or test situation.

While most of the items were ranked differently as a result of the index used, the differences were usually not large enough to eliminate an item when the best two-thirds were compared. However, in most cases where CR tests are used, information is needed which is usually very curriculum-specific. Test results more often than not are used as diagnostic measures to identify specific skills or

lack of such skills. Items which may be eliminated by the traditional point biserial, as few as they may be, may be covering important objectives and may be especially valuable for diagnostic purposes. It is very possible that an item which is discriminating well between students needing training and those not needing training in a specific skill is the same item eliminated by the traditional point biserial. The fewer the items measuring a given objective, the greater the effect of this error.

The correlations between the three indices did not seem to be related to observed score variance in any consistent way. All samples with the exception of the low achievement reading component yielded significant correlations between indices. These correlations were somewhat higher for the low and unrestricted samples than for the high and middle achievement samples. The low achievement; reading component seemed to be the only sample and test situation which caused the relationship between indices to depart from the relationship exhibited by the other achievement samples. For this data, correlations were found between r_t vs. ϕ , and r_t vs. r_e which were quite lower than those obtained from the other samples. However, for the same data, a high correlation was obtained between ϕ and r_e . It would seem that the traditional point biserial (r_t), when compared to either of the two criterion-referenced techniques, yields a different ordering of items for this sample and test situation only.

Reliability

For each set of items selected by the three indices, Livingston's reliability coefficient, Harris' index of efficiency, and the traditional KR-20 coefficient were computed. Table 6 presents these data for each sample and test situation.

A comparison of the absolute values of the three indices by sample and test situation shows that Harris' index was consistently the lowest while Livingston's was consistently the highest. In a usual test evaluation context, these absolute values would be considered important and would be used to indicate the decision-making accuracy of the instrument. A large value would indicate a high decision-making accuracy. While comparisons between samples and test situations for a given index is legitimate, comparisons between indices, in terms of absolute values, for the same sample and test situation are somewhat

inappropriate due to the statistical nature of the indices. For example, Harris' index is very sensitive to the number of masters vs. non-masters. The greater the proportion of masters, the smaller the index will tend to be. However, even though Harris' index may be relatively small, the decisions based on it may be substantially accurate (Hambleton, 1975). The present data consists of relatively high proportions of students reaching the criterion and being judged masters. The low achievement sample, which by definition contains the smallest number of students reaching mastery, has from 84 to 88 percent reaching mastery depending on item selection and test situation. This sample also yielded the highest indices, .46 to .51.

Since Livingston's coefficient redefines the variance as a deviation from the mastery criterion score rather than from the mean, this coefficient will be at least as large as KR-20. Because of this, the appearance of large coefficients is not unexpected. The greater the difference between the mean and the cut-off score, the greater the reliability coefficient will be.

The results of this study indicate that KR-20 coefficients provide a more consistent estimate of reliability for CR tests than either of the other two coefficients. Livingston's and Harris' coefficients seem to be inconsistent with regard to both observed score variability and item homogeneity. For both coefficients, it seems likely that a value obtained for one group may not have the same meaning as a value obtained from a second more homogeneous group. KR-20 estimates did not reflect these problems. KR-20 coefficients seem to remain consistent despite the manipulations in both observed score variance and item homogeneity.

As was expected, KR-20 was directly related to sample variance. There was a corresponding increase in KR-20 as subject variance increased regardless of test situation. No direct relationship could be noted between the other two indices and subject variance, or between one another. This is not to say that they were not responding to the change in subject variance; they were simply not responding in any consistent manner which would allow an interpretation to be made. Interpretations based on Livingston's and Harris' coefficients were difficult, if not impossible. At best, it would seem that different interpretations would be required for different testing situations.

There seems to be a correspondence between increase in test homogeneity and increases in reliability estimates for all three indices. However, this relationship is far more evident with KR-20 than with either of the other two indices. This is revealed by both the magnitude of the increases for KR-20 as well as the consistency across achievement samples.

Many researchers (Shavelson, Block, and Ravitch, 1972) have questioned the use of traditional internal consistency methods such as KR-20 for CR tests which contain several scales. Their feeling is that the test would not be highly internally consistent and as a result, such indices would be inappropriate. The findings of the present study do not support this suspicion. KR-20 is a consistent estimate of reliability even when multiscale CR tests are being considered. If all scales of a multiscale CR test are highly intercorrelated and have a commonality with the domain, the fullscale homogeneity may be reasonably high.

Conclusions

The results of the present study indicate that the construction of CR tests can be greatly facilitated by item analysis procedures such as phi, and the application of traditional reliability estimates such as KR-20. The procedure of trying out test items on a group of examinees consisting of master and non-masters is recommended.

References

- Haladyna, T.M. Effects of different samples on item and test characteristics of criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 93-99
- Hambleton, R.K., Swaminathan, H., Algina, J. & Coulson, D. Criterion-referenced testing and measurement: a review of technical issues and developments. An invited symposium presented at the meeting of the American Educational Research Association, Washington, D.C., April 1975
- Klein, S.P., & Kosecoff, J. Issues and procedures in the development of criterion-referenced tests. ERIC TM Report 26, September, 1973.
- Millman, J., & Popham, W.J. The issue of item and test variance for criterion-referenced tests: A clarification. Journal of Educational Measurement, 1974, 11, 137-138.
- Popham, W.J. Measurement advances for educational evaluators. Paper presented at the meeting of the American Educational Research Association, Chicago, April 1974.
- Popham, W.J. & Husek, T.R. Implications of criterion-referenced measurement. Journal of Educational Measurement, 1969, 6, 1-9.
- Rentz, R.R. Technical Manual: Language Skills Examination. Board of Regents, University System of Georgia, Atlanta, Ga., 1971.
- Roudabush, G.E. Item selection of criterion referenced tests. Paper presented at the meeting of the American Educational Research Association, New Orleans, 1973.
- Shavelson, R., Block, J., & Ravitch, M. Criterion-referenced testing: comments on reliability. Journal of Educational Measurement, 1972, 2, 133-138.
- Woodson, M.I.C.E. The issue of item and test variance for criterion-referenced tests. Journal of Educational Measurement, 1974, 11, 63-64.

Sample and Test Situation	Condition	Item Selection Method	Reliability
<u>Total Sample by Total Test</u>			
unrestricted sample, reading ... and writing combined	high subject variability low item homogeneity	r_t $\emptyset$ r_e	KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2
<u>Total Sample by Reading Component</u>			
unrestricted sample, reading component	high subject variability high item homogeneity	r_t $\emptyset$ r_e	KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2
<u>Sub-Samples by Total Test</u>			
low achievement sample, writing and reading combined	low subject variability low item homogeneity	r_t $\emptyset$ r_e	KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2
middle achievement sample, writing and reading combined	low subject variability low item homogeneity	r_t $\emptyset$ r_e	KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2
high achievement sample, writing and reading combined	low subject variability low item homogeneity	r_t $\emptyset$ r_e	KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2 KR-20 r_{cc} uc^2

Figure 1
Summary of Comparisons to Be Used in Analysis

Sample and Test Situation	Condition	Item Selection Method	Reliability
<u>Sub-Samples by Reading Component</u>			
low achievement sample, reading component	low subject variability high item homogeneity	r_t	KR-20
		ϕ	r_{cc}
		r_e	uc2
middle achievement sample, reading component	low subject variability high item homogeneity	r_t	KR-20
		ϕ	r_{cc}
		r_e	uc2
high achievement sample, reading component	low subject variability high item homogeneity	r_t	KR-20
		ϕ	r_{cc}
		r_e	uc2

r_t = Point biserial coefficient using total test score as criterion.
 ϕ = Phi coefficient, essay test determines mastery and non-mastery.
 r_e = Point biserial coefficient using essay test score as criterion.
 KR-20 = Kuder-Richardson.
 r_{cc} = Livingston's reliability coefficient.
 uc2 = Harris' index of efficiency.

Figure 1 Continued

Summary of Comparisons to Be Used in Analysis

Table 1

Score Categories, Per Cent in Stanine and Number of Subjects
Used to Create Restricted and Unrestricted Distributions

Score Category on Writing Component	Restricted Distributions			Unrestricted Distributions		
	Low Achievement	Middle Achievement	Low Achievement	Percent in Stanine	Percent in Stanine	Across Achievement
	Percent in Stanine	Percent in Stanine	Percent in Stanine	Percent in Stanine	Percent in Stanine	Percent in Stanine
	N	N	N	N	N	N
11	4			4	16	
12	7			7	28	
13	12			12	48	
14	17			17	58	
15	20	4	16	20	80	
16	17	7	28	17	68	
17	12	12	48	12	48	
18	7	17	68	7	28	
19	4	20	80	4	16	
20		17	68		28	
21		12	48		48	
22		7	28		58	
23		4	16		80	
24					68	
25					48	
26					28	
27					16	
Totals	100%	100%	100%	100%	100%	400

TABLE 2
t TEST FOR MEANS AND F RATIOS FOR
VARIANCE BETWEEN ALL ACHIEVEMENT SAMPLES

SAMPLES	READING COMPONENT		WRITING AND READING COMPONENT	
	t	F	t	F
Low Achievement vs. Middle Achievement	6.72**	1.07	11.33**	1.05
Low Achievement vs. High Achievement	15.15**	1.38**	26.13**	1.54**
Low Achievement vs. Unrestricted	6.58**	1.07	11.02**	1.31**
Middle Achievement vs. High Achievement	8.19**	1.29 *	13.05**	1.70**
Middle Achievement vs. Unrestricted	0.08	1.14	0.24	1.19
High Achievement vs. Unrestricted	7.80**	1.46**	10.87**	2.01**

* $p \leq .05$
 ** $p \leq .01$

TABLE 3

SPEARMAN RANK ORDER CORRELATIONS BETWEEN
DISCRIMINATION INDICES

	r_t vs ρ	r_t vs r_e	ρ vs r_e
Unrestricted Sample with: Writing and reading combined	.74*	.75*	.93*
Reading component	.53*	.62*	.89*
High Achievement Sample With: Writing and reading combined	.65*	.72*	.81*
Reading component	.63*	.63*	.78*
Middle Achievement Sample With: Writing and reading combined	.68*	.71*	.94*
Reading component	.56*	.61*	.87*
Low Achievement Sample With: Writing and reading combined	.76*	.78*	.99*
Reading component	.27	.33	.96*

*p < .01

TABLE 4

PERCENTAGE OF ITEMS SELECTED IN COMMON FOR
DIFFERENT TEST LENGTHS BY SAMPLE AND TEST SITUATION

Sample and Test Situation	Final Test Length as Portion of Item Pool									
	Top One Third			Top Two Thirds			Bottom One Third			
	r_t vs $\emptyset$	r_t vs r_a	$\emptyset$ vs r_e	r_t vs $\emptyset$	r_t vs r_a	$\emptyset$ vs r_a	r_t vs $\emptyset$	r_t vs r_a	$\emptyset$ vs r_a	
Unrestricted Sample Reading Component Only	46%	62%	77%	82%	82%	93%	52%	62%	35%	
High Achievement Sample Reading Component Only	62%	76%	69%	78%	78%	82%	54%	54%	52%	
Middle Achievement Sample Reading Component Only	39%	62%	69%	85%	85%	93%	69%	59%	35%	
Low Achievement Sample Reading Component Only	31%	39%	85%	74%	78%	96%	46%	54%	92%	
Unrestricted Sample Writing and Reading Combined	69%	69%	85%	89%	89%	96%	77%	77%	92%	
High Achievement Sample Writing and Reading Combined	77%	69%	85%	89%	85%	82%	59%	52%	52%	
Middle Achievement Sample Writing and Reading Combined	54%	54%	92%	85%	85%	89%	69%	69%	77%	
Low Achievement Sample Writing and Reading Combined	69%	69%	92%	89%	89%	100%	77%	77%	100%	

TABLE 5

MEANS, STANDARD DEVIATIONS AND RELIABILITIES
FOR SAMPLES AND TEST SITUATIONS BY
ITEM SELECTION METHOD

SAMPLE AND TEST SITUATION	ITEM SELECTION METHOD									
	r_t		θ		r_e		Range			
	Mean	S.D.	Mean	S.D.	Mean	S.D.	Mean	KR-20	Mean	KR-20
Unrestricted sample										
reading and writing										
combined:										
reading component	18.57	4.05	18.93	3.86	18.58	3.85	.37	.03		
alone:	18.71	4.98	18.51	4.76	18.45	4.76	.27	.04		
High Achievement										
sample										
reading and writing										
combined:	19.68	3.10	19.95	2.81	19.81	2.89	.28	.07		
reading component	20.08	4.36	20.55	4.03	19.41	4.09	1.15	.04		
alone:										
Middle Achievement										
sample										
reading and writing										
combined:	18.79	3.48	19.24	3.33	18.42	3.36	.83	.04		
reading component	18.21	4.76	18.43	4.60	18.12	4.58	.32	.03		
alone:										
Low Achievement										
sample										
reading and writing										
combined:	16.54	3.78	16.84	3.69	16.84	3.69	.31	.02		
reading component	17.64	4.80	17.51	4.24	17.65	4.35	.15	.08		
alone:										
							.46	.04		
						MEAN				

Table 6
MEANS, STANDARD DEVIATIONS, AND RELIABILITIES
BY SAMPLE AND TEST SITUATIONS FOR TESTS
SELECTED BY THE THREE INDICES

UNRESTRICTED SAMPLE WITH READING AND WRITING COMBINED AND
UNRESTRICTED SAMPLE WITH READING COMPONENT

UNRESTRICTED SAMPLE WITH READING AND WRITING COMBINED							
Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion $\frac{N}{\bar{x}}$	KR-20	$\frac{uc}{2}$	r_{cc}	
r_t	18.57	4.05	367	.73	.36	.91	
$\emptyset$	18.93	3.86	374	.71	.34	.91	
r_e	18.58	3.85	373	.71	.33	.91	
UNRESTRICTED SAMPLE WITH READING COMPONENT							
Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion $\frac{N}{\bar{x}}$	KR-20	$\frac{uc}{2}$	r_{cc}	
r_t	18.71	4.98	350	.83	.45	.93	
$\emptyset$	18.51	4.76	355	.80	.41	.92	
r_e	18.45	4.55	356	.79	.41	.92	

Table 6 (Con't)

MEANS, STANDARD DEVIATIONS, AND RELIABILITIES
BY SAMPLE AND TEST SITUATIONS FOR TESTS
SELECTED BY THE THREE INDICES

HIGH ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED AND
HIGH ACHIEVEMENT SAMPLE WITH READING COMPONENT

HIGH ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED							
Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion	KR-20	uc^2	r_{cc}	
			$\frac{N}{\bar{x}}$				
rt	19.68	3.10	394	.57	.11	.92	
ø	19.95	2.81	398	.51	.05	.93	
re	19.81	2.89	398	.54	.04	.93	

HIGH ACHIEVEMENT SAMPLE WITH READING COMPONENT

HIGH ACHIEVEMENT SAMPLE WITH READING COMPONENT							
Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion	KR-20	uc^2	r_{cc}	
			$\frac{N}{\bar{x}}$				
rt	20.08	4.36	371	.79	.38	.94	
ø	20.55	4.03	380	.76	.29	.95	
re	19.41	4.09	366	.76	.41	.93	

Table 6 (Cont)

MEANS, STANDARD DEVIATIONS, AND RELIABILITIES
BY SAMPLE AND TEST SITUATIONS FOR TESTS
SELECTED BY THE THREE INDICES

MIDDLE ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED AND
MIDDLE ACHIEVEMENT SAMPLE WITH READING COMPONENT

MIDDLE ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED

Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion		KR-20	$\overline{uc^2}$	r_{cc}
			N	%			
rt Ø re	13.79	3.48	379	95	.63	.27	.90
	19.24	3.33	383	96	.60	.26	.91
	18.42	3.36	377	94	.62	.29	.89

MIDDLE ACHIEVEMENT SAMPLE WITH READING COMPONENT

Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion		KR-20	$\overline{uc^2}$	r_{cc}
			N	%			
rt Ø re	18.21	4.76	343	86	.80	.50	.91
	18.43	4.60	353	88	.78	.45	.91
	18.12	4.58	344	86	.78	.49	.90

Table 6 (Con't)

MEANS, STANDARD DEVIATIONS, AND RELIABILITIES
BY SAMPLE AND TEST SITUATIONS FOR TESTS
SELECTED BY THE THREE INDICES

LOW ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED AND
LOW ACHIEVEMENT SAMPLE WITH READING COMPONENT

LOW ACHIEVEMENT SAMPLE WITH READING AND WRITING COMBINED

Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion	KR-20	$\overline{uc^2}$	$\overline{r_{cc}}$
			$\frac{N}{\%}$			
r_t	16.54	3.78	334	.64	.51	.81
$\emptyset$	16.84	3.69	347	.63	.46	.82
r_e	16.84	3.69	347	.63	.46	.82

27

LOW ACHIEVEMENT SAMPLE WITH READING COMPONENT

Item Selection Method	Test Mean	S.D.	Number of Subject Reaching Criterion	KR-20	$\overline{uc^2}$	$\overline{r_{cc}}$
			$\frac{N}{\%}$			
r_t	17.64	4.80	338	.80	.50	.90
$\emptyset$	17.51	4.24	353	.73	.46	.87
r_e	17.65	4.35	352	.74	.46	.88